Towards Unified Representation of Invariant-Specific Features in Missing Modality Face Anti-Spoofing

Guanghao Zheng[1†], Yuchen Liu[2†], Wenrui Dai[1⋆], Chenglin Li[2] and Junni Zou[1], and Hongkai Xiong[2]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] Department of Electronic Engineering, Shanghai Jiao Tong University
{zgh990318,liuyuchen6666, daiwenrui, lcl1985, zoujunni, xionghongkai}@sjtu.edu.cn

A. Proof of Proposition 1

Proposition 1 *The LBP-Guided Contrastive Loss promotes to learn modality-specific features by increasing the mutual information $I(X;X^+)$ but decreasing $I(X;X^-)$, where X, X^+ and X^- are random variables of anchor, positive samples and negative samples.*

Proof. We prove that the LBP-guided contrastive loss can increase the mutual information $I(X;X^+)$ but decrease $I(X;X-)$, where anchor, positive, and negative samples, x, x^+, x^- are drawn from the random variables X, X^+, X^-.

$$
\begin{aligned}
&I(X;X^+) = \mathbb{E}_{p(x,x^+)}\left[\log\frac{q(x|x^+)}{p(x)}\right] + \mathbb{E}_{p(x^+)}[KL(p(x|x^+)\|q(x|x^+))] \\
&\geqslant \mathbb{E}_{p(x,x^+)}\left[\log\frac{q(x|x^+)}{p(x)}\right] \\
&\approx \mathbb{E}\left[\log\frac{\exp([\mathrm{sim}(x_i,x_i^+)+\mathrm{sim}(AT(x_i),AT(x_i^+))]/\tau)}{\frac{1}{2K-1}\sum_{i=1}^{K}(\exp(\mathrm{sim}(x_i,x_i^+)/\tau)+\exp(\mathrm{sim}(x_i,x_i^-)/\tau))}\right] \\
&= \log(2K-1) - L_{con} \\
&q(x|x^+) = p(x)\frac{\exp([\mathrm{sim}(x,x^+)+\mathrm{sim}(AT(x),AT(x^+))]/\tau)}{\mathbb{E}_{p(x^+)}[\exp([\mathrm{sim}(x,x^+)+\mathrm{sim}(AT(x),AT(x^+))]/\tau]}
\end{aligned}
$$

Here, K is the batch size and AT is the attack type considered to measure the similarity of two samples in FAS. $\mathrm{sim}(AT(x), AT(x^+))$ is measured using LBP feature similarity. We prove that minimizing L_{con} increases the mutual information $I(X;X^+)$, and similarly, minimizing L_{con} decreases $I(X;X^-)$. Thus, higher $I(X;X^+)$ and lower $I(X;X^-)$ imply less mutual information across different modal combinations and suggest learning modality-specific features clustered for each modal combination.

⋆ Correspondence to Wenrui Dai. † Equal contribution.

Table S-1: A summary of the FAS datasets used in our experiments.

Datasets	Year	Subjects	Videos	Modal Types	Attack Types
WMCA [2]	2019	72	1,941	RGB/Depth/IR/Thermal	2 Print, Replay, 2D/3D Mask
CASIA-SURF [6]	2019	1,000	21,000	RGB/Depth/IR	Print, Cut
CASIA-SURF-CeFA [3]	2019	1,607	23,538	RGB/Depth/IR	Print, Replay,3D mask

B. Detailed Explanations

The diagram of LBP-Guided Contrastive Loss.

We explain the diagram details in Fig. 3. Taking sample #1 as an anchor, samples #2 and #3 with batch-level masking are positive samples as they have the same modal combination and label with the anchor. Sample #4 is negative sample with different modal combination and label. The red opposite arrow means the anchor and positive samples are pulled together, while the red two-way arrow means the anchor and negative samples are pushed away. Thus, in order to enhance the modality-specific features, the LBP-Guided Contrastive Loss pull in samples with the same modal combination, while pushing away samples with different modal combinations.

The reason for introducing batch samples.

Batch samples are defined as samples having the same modal combination with the anchor in Eq. (3), which are used to avoid the situation of no positive samples. To be specific, sample-level masking randomly masks samples to seven modal combinations. Thus, samples with the same label might have different modal combinations from the anchor, and no positive samples are obtained. In contrast, batch-level masking enforces all the samples in a batch to be masked with the same modal combination, which guarantees the existence of samples with the same label and modal combination as positive samples.

C. Datasets

Experiments are conducted on the three publicly available datasets: WMCA [2], CASIA-SURF [6] and CASIA-SURF-CeFA [3]. These datasets contain three modalities (RGB,Depth and IR) for spoofing recognition. Basic information of these datasets is summarized in Table S-1.

- **WMCA** consists of 1941 short video recordings of both bonafide and presentation attacks from 72 different identities. The data is recorded from several channels including color, depth, infra-red, and thermal.
- **CASIA-SURF** is a large-scale multi-modal face anti-spoofing dataset with 1, 000 subjects and 3 modalities. It only consists two 2D attack types, which is simpler than the other two datasets.
- **CASIA-SURF-CeFA** covers 3 ethnicities, 3 modalities, 1, 607 subjects, and 2D plus 3D attack types.

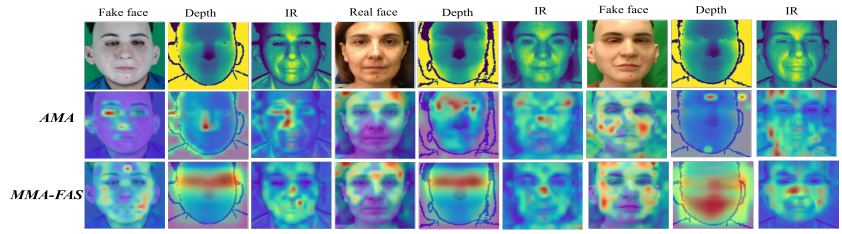

Fig. S-1: Attention maps in MultiViT backbone when applying AMA and MMA-FAS.

D. Implementation Details

We leverage MultiViT-B [1] as our backbone, which consists 3 tokenizers and 12 transformer blocks. The prediction head is replaced by a binary classification head. We apply DeiT-B [4] pretrained weight for initialization of MultiViT-B. For tokenizers, the number and dimension of patch embeddings are $D = 768$ and $N_t = 589$. For AMA, the kernel size of the filtering mask in MDA f is 2. The trade-off parameter λ is set to 0.1. We use Adam optimizer and the learning rate is set to 7e-4 with a cosine decline schedule. We train 100 epochs with batch size 32 and weight decay 5e-3. We leverage 2 NVIDIA GTX- 2080Ti GPUs for training.

Table S-2: ACER on WMCA dataset with different blocks frozen. 'CH' denotes that only classification head is finetuned.

Methods	Modalities						
	R	D	I	R+D	R+I	D+I	R+D+I
All blocks	22.04	14.95	24.06	12.69	20.44	15.23	10.47
Last block	5.76	5.28	3.55	3.64	2.32	2.78	2.28
Last 2 block	8.26	6.77	6.80	5.48	4.59	3.73	2.54
Last 4 block	6.68	6.22	7.59	6.01	5.84	4.25	3.95
CH	**4.42**	**3.63**	**1.46**	**1.80**	**1.38**	**1.88**	**1.25**

Table S-3: ACER on the WMCA dataset with different forward.

Methods	Batch Size	Modalities						
		R	D	I	R+D	R+I	D+I	R+D+I
One Forward	16	5.28	7.66	4.21	4.34	3.75	3.31	2.96
Two Forward	8	**4.42**	**3.63**	**1.46**	**1.80**	**1.38**	**1.88**	**1.25**

Table S-4: ACER on cross-dataset and cross-attack protocol.

Methods	Modalities							
	R	D	I	R+D	R+I	D+I	R+D+I	Avg.
Vanilla ViT	51.09	50.03	50.00	49.87	**49.62**	50.30	43.23	49.16
AMA [20]	49.27	50.22	50.00	50.06	50.00	50.22	**36.85**	48.08
MAP [8]	58.67	51.22	51.54	50.46	49.97	49.68	41.13	57.59
MMA-FAS	**48.59**	**45.61**	**48.82**	**44.13**	49.87	**42.26**	37.22	**45.21**

E. Extensive Experiments

Different frozen blocks. Several transformer block in MultiViT-B [1] are frozen during training, since training all the blocks leads to the overfitting problem in multi-modal FAS. Table S-2 compares ACER by freezing different blocks of the backbone and shows that finetuning only the classification head achieves the best performance while finetuning all blocks the worst.

Experiments on cross-dataset and cross-attack type protocol. We further evaluate our method on more complex protocol.In this protocol, models are trained on CASIA-SURF dataset and tested on WMCA dataset ('flexible-mask' protocol). Table S-4 shows MMA-FAS outperforms other methods greatly (at least 2.85% lower ACER).

Comparisons between one or two forward propagation. Our MMA-FAS needs two forward propagation in the training stage. However, using a large batch size achieves batch-level and sample-level masking in one forward propagation. In this case, positive and negative samples have to be masked from different samples, which introduces sample bias during learning. In contrast, we mask the same sample into two modal combinations to serve as positive samples and negative samples by separately using batch-level and sample-level masking in two forward propagation. Contrasting anchor samples with these positive and negative samples eliminates sample bias and promotes the learning of modality-specific features. Table S-3 shows that two forward propagation yield better performance.

Visualizations. Fig. S-1 visualizes the attention maps in MultiViT backbones in WMCA dataset. Different from the AMA [5] which learns modality-invariant features and focuses on similar regions in three modalities, MMA-FAS extracts both modality-specific and modality-invariant features simultaneously. For example, the MultiViT in MMA-FAS focuses on cheeks and forehead in RGB images, eye region in Depth images, and nose and mouth in IR images.

References

1. Bachmann, R., Mizrahi, D., Atanov, A., Zamir, A.: Multimae: Multi-modal multi-task masked autoencoders. In: European Conference on Computer Vision. pp. 348–367. Springer (2022)
2. George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., Marcel, S.: Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE transactions on information forensics and security **15**, 42–55 (2019)
3. Liu, A., Tan, Z., Wan, J., Escalera, S., Guo, G., Li, S.Z.: Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1179–1187 (2021)
4. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers and distillation through attention (2020). doi: 10.48550. arxiv (2012)

5. Yu, Z., Cai, R., Cui, Y., Liu, X., Hu, Y., Kot, A.: Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. arXiv preprint arXiv:2302.05744 (2023)
6. Zhang, S., Liu, A., Wan, J., Liang, Y., Guo, G., Escalera, S., Escalante, H.J., Li, S.Z.: Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. IEEE Transactions on Biometrics, Behavior, and Identity Science **2**(2), 182–193 (2020)